



AFRL-OSR-VA-TR-2013-0062

NON-PARAMETRIC BAYESIAN ANALYSIS OF HETEROGENEOUS DATA

David Blei

Princeton University

March 2013

Final Report

DISTRIBUTION A: Approved for public release.

**AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND**

REPORT DOCUMENTATION PAGE				<i>Form Approved</i> <i>OMB No. 0704-0188</i>	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Services and Communications Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.					
1. REPORT DATE (DD-MM-YYYY) 1/7/2013		2. REPORT TYPE FINAL		3. DATES COVERED (From - To) 9/15/2009-9/14/2012	
4. TITLE AND SUBTITLE NON-PARAMETRIC BAYESIAN ANALYSIS OF HETEROGENEOUS DATA				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER FA9550-09-1-0668	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) David Blei				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Princeton University 4 New South Building Princeton, NJ 08544				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR 875 N Randolph St Arlington, VA 22203 Dr. Tristan Nguyen/RSL				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-OSR-VA-TR-2013-0062	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution A: Approved for Public Release					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Under this grant, my research focused on fusing heterogenous sources of data with Bayesian nonparametric models. We published many papers in the service of this goal. I would like to highlight the following papers about furthering Bayesian nonparametrics and examining the fusion of heterogenous data types in a diversity of settings. This is an extension of last year's report. It is my final report.					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT U	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Jeffrey Friedland
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (Include area code) 609-258-6325

Reset

Report for AFOSR 09NL202

David M. Blei
Princeton University

December 18, 2012

Under this grant, my research focused on fusing heterogeneous sources of data with Bayesian nonparametric models. We published many papers in the service of this goal. I would like to highlight the following papers about furthering Bayesian nonparametrics and examining the fusion of heterogeneous data types in a diversity of settings. This is an extension of last year's report. It is my final report.

1. With Sam Gershman, we wrote a tutorial about Bayesian nonparametrics (Gershman and Blei, 2012).
2. With Peter Frazier and colleagues, we have worked on *distance dependent* Bayesian nonparametric models (Blei and Frazier, 2011; Gershman et al., 2011; Ghosh et al., 2011). These allow external data sources to influence the latent clustering (and latent feature representation) of a variety of data. We have applied these models to text, images, EEG, and stock prices.
3. With Lauren Hannah, we developed *Dirichlet process mixtures of generalized linear models* (Hannah et al., 2010, 2011). These allow covariates to affect the clustering of a response and exert a relationship on it.
4. With Chong Wang, we modeled collaborative filtering data—user preferences and *content* about the items (Wang and Blei, 2011). This work won the **Best Student Paper Award** at KDD 2011.
5. With Sean Gerrish, we built a model of legislative roll call data (i.e., votes on bills) and bill texts (Gerrish and Blei, 2011). This work won a **Distinguished Application Award** at ICML 2011. We recently furthered this work to model issue-adjusted ideal points (Gerrish and Blei, 2012).
6. John Paisley, Chong Wang, and I developed the *Discrete Infinite Logistic Normal* (DILN), which is a new kind of Bayesian nonparametric model (Paisley et al., 2011, 2012). DILN allows the atoms of an underlying random measure to exert correlation.

7. To perform inference with massive data sets, Matt Hoffman, Francis Bach, and I developed stochastic variational inference for Latent Dirichlet allocation (Hoffman et al., 2010a). Chong Wang, John Paisley, and I extended this algorithm to the hierarchical Dirichlet process, enabling us to fit Bayesian nonparametric models to massive data (Wang et al., 2011). Recently, Chong Wang and I developed a truncation-free variant of stochastic variational inference for this important class of models (Wang and Blei, 2012).
8. Jonathan Chang and I published the *relational topic model*, a model of documents and links (Chang and Blei, 2010). Unlike traditional network models, this model incorporates node content—it can predict content from links and links from content. Prem Gopalan and I developed stochastic inference for analyzing massive social networks (Gopalan et al., 2012).
9. Matt Hoffman and I wrote several papers about Bayesian nonparametric analysis of recorded music (Hoffman et al., 2009b,a,c, 2010b).
10. Chong Wang and I developed a variational inference algorithm for the nested Chinese restaurant process (Wang and Blei, 2009b).
11. Chong Wang and I relaxed some of the assumptions made by the hierarchical Dirichlet process, coupling sparsity and smoothness (Wang and Blei, 2009a). With Sinead Williamson and Katherine Heller, we further extended this work to matrix factorization (Williamson et al., 2010).

References

- Blei, D. and Frazier, P. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.
- Chang, J. and Blei, D. (2010). Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1).
- Gerrish, S. and Blei, D. (2011). Predicting legislative roll calls from text. In *International Conference on Machine Learning*.
- Gerrish, S. and Blei, D. (2012). How they vote: Issue-adjusted models of legislative behavior. In *Neural Information Processing Systems*.
- Gershman, S. and Blei, D. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12.
- Gershman, S., Frazier, P., and Blei, D. (2011). The distance-dependent Indian buffet process. *Journal of the American Statistical Association* (submitted).
- Ghosh, S., Ungureanu, A., Sudderth, E., and Blei, D. (2011). Spatial distance dependent Chinese restaurant processes for image segmentation. In *Neural Information Processing Systems*.

- Gopalan, P., Mimno, D., Gerrish, S., Freedman, M., and Blei, D. (2012). Scalable inference of overlapping communities. In *Neural Information Processing*.
- Hannah, L., Blei, D., and Powell, W. (2010). Dirichlet process mixtures of generalized linear models. In *Artificial Intelligence and Statistics*.
- Hannah, L., Blei, D., and Powell, W. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, to appear.
- Hoffman, M., Blei, D., and Bach, F. (2010a). On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems*.
- Hoffman, M., Blei, D., and Cook, P. (2009a). Easy as CBA: A simple probabilistic model for tagging music. In *International Conference on Music Information Retrieval*.
- Hoffman, M., Blei, D., and Cook, P. (2009b). Finding latent sources in recorded music with a shift-invariant HDP. In *International Conference on Digital Audio Effects*.
- Hoffman, M., Blei, D., and Cook, P. (2010b). Bayesian nonparametric matrix factorization for recorded music. In *International Conference on Machine Learning*.
- Hoffman, M., Cook, P., and Blei, D. (2009c). Bayesian spectral matching: Turning young MC into MC hammer via MCMC sampling. In *International Computer Music Conference*.
- Paisley, J., Wang, C., and Blei, D. (2011). The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*.
- Paisley, J., Wang, C., and Blei, D. (2012). The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(2):235–272.
- Wang, C. and Blei, D. (2009a). Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1982–1989.
- Wang, C. and Blei, D. (2009b). Variational inference for the nested Chinese restaurant process. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1990–1998.
- Wang, C. and Blei, D. (2011). Collaborative topic modeling for recommending scientific articles. In *Knowledge Discovery and Data Mining*.
- Wang, C. and Blei, D. (2012). Truncation-free stochastic variational inference for Bayesian nonparametric models. In *Neural Information Processing*.
- Wang, C., Paisley, J., and Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*.
- Williamson, S., Wang, C., Heller, K., and Blei, D. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *International Conference on Machine Learning*.